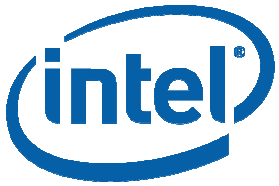
IA-64 Kexec/Kdump

Zou Nan hai

Open Source Technology Center



Kexec/Kdump

Open Source Technology Center

- Kexec is a feature for Linux* to allow a running Linux kernel to load and run another Linux* kernel.
- Kdump is a kexec based kernel crash dumping mechanism.

intel

Kexec

Open Source Technology Center

- Bypass hardware reset
- Bypass firmware.
- Bypass boot loader
- Speed up reboot time
- May used to support Kdump.

intel

Using Kexec

Open Source Technology Center

- Kexec has 2 components

  a Kexec-tools and some kernel patches. The kernel patches are already in mainstream kernel.

- Build and reboot a kernel with configure option CONFIG_KEXEC=y
- Build and install kexec-tools.
- Load the new kernel with command

  “kexec –l <kernel-image> --append=<command line options> --initrd=<initrd>”

- Boot to the new kernel with command

  “kexec –e”

intel

Kexec

Open Source Technology Center

- When executing “kexec –l”,
  kexec-tools will interpret the content of kernel image file.
  Load the segments together with initrd image and command line to kernel space, via a system call sys_kexec_load().
  Kexec-tools will determine the layout of those segment with the help of /proc/iomem.

- Kernel use a structure of kexec_segment to track the loaded segment.

intel

Kexec

Open Source Technology Center

- On some platforms, the kernel is build at fixed physical address, the first kernel and the second kernel will overlap, so do memory copy in the first kernel itself may not be safe.
- Kexec will allocate one or few control pages to contain a small piece of assembly code called relocate code.

  Kernel will make sure control pages will not overlap with neither of the first kernel nor the second kernel, thus perform page copy in relocate code is safe.
- Kexec –e will call sys_reboot with a special flag LINUX_REBOOT_CMD_KEXEC

intel

Kexec

Open Source Technology Center

- Kernel travels the device list, call device driver's shutdown method on each device.
- Kernel will then jump to relocate code via machine_kexec(), relocate code will do some architecture depended low level work. copy the segments to their destination.

  then jumps to the new kernel.

intel

Kexec on IA-64

Open Source Technology Center

- In machine kexec, put all the other CPUS to SAL rendez state with the help of CPU hotplug.
- Clear possible pending interrupts.
- Translate to physical mode first to avoid later TLB translation.
- Purge all TC and TR entries.
- Copy segments to their destination, flush icache.
- Prepare stack and initialize register stack.
- Jump to the new entry.

intel

Purgatory code

Open Source Technology Center

- The new entry is not the entry point of the new kernel, but the entry of the purgatory code.
- Purgatory code is some pre compiled code in kexec-tools. It is designed to simplify kernel code.
- It is compiled at the time of kexec-tools is build.
- It is loaded as one of the kexec segment.
- It is linked by kexec-tools when “kexec –l” is executed, after memory layout is determined.
- It is executed between the first and the second kernel.

intel

Purgatory code (IA-64)

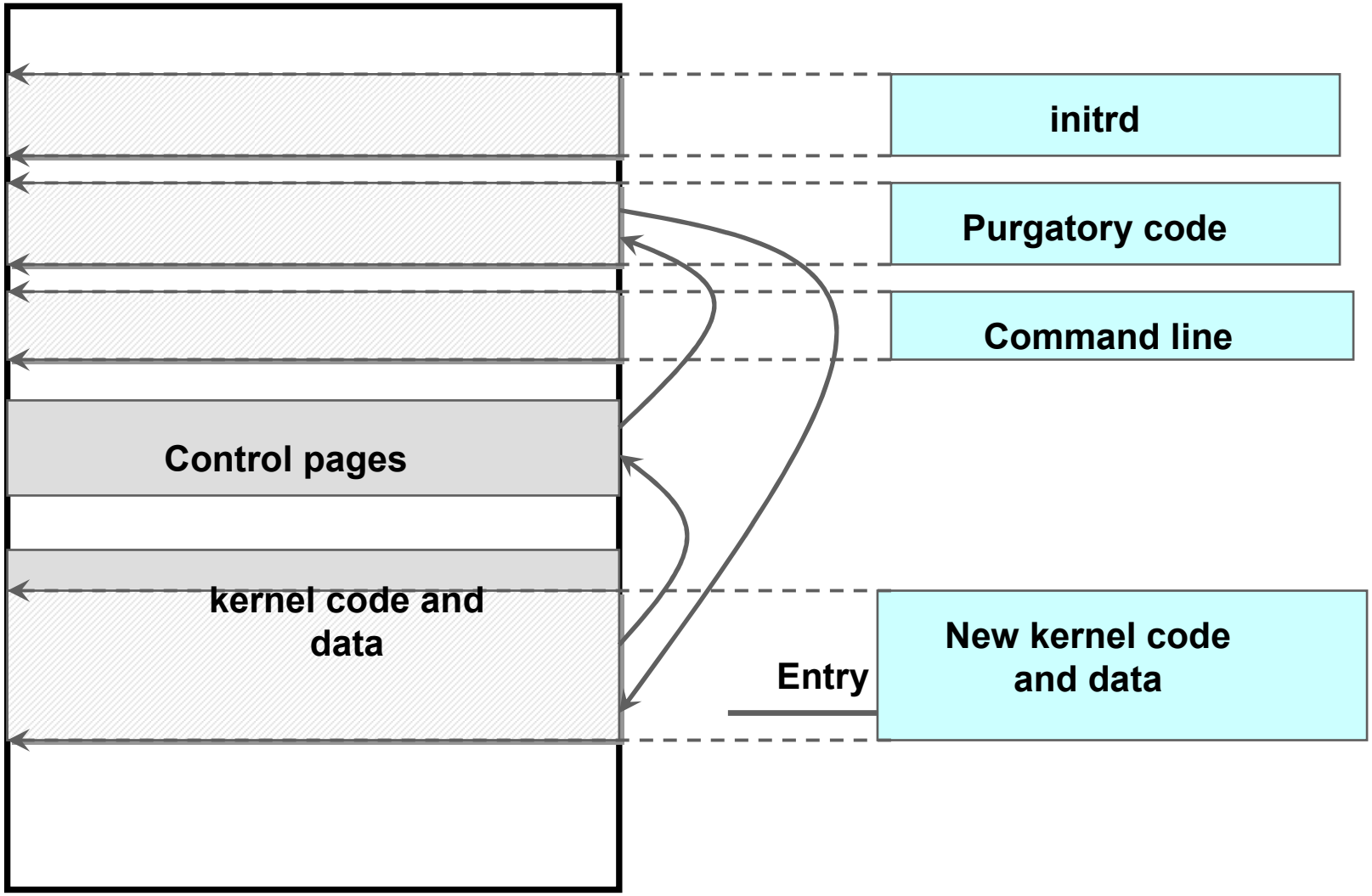
Open Source Technology Center

- Prepare boot_parameter(command line, initrd position etc)
- Hook an empty set_virtual_address_map to EFI runtime, necessary to kexec to an unmodified kernel.
- Patch EFI memmap if it is kdump.
- Boot to the real new kernel entry with the prepared boot_parameter in register r28.

intel

Kexec

Open Source Technology Center



intel

Kdump

Open Source Technology Center

- Kdump is a Kexec based feature on Linux* to capture system crash dump.
- Reserve a region for dump capture kernel.
- Kexec to dump capture kernel when kernel panic, machine check or INIT even was triggered.
- Dump capture kernel can access memory of crashed kernel via /proc/vmcore.
- Post analysis kernel crash with vmcore file with crash-utility or gdb.

intel

Crash dump solutions

Open Source Technology Center

- Disk Dump
  - dump to disk via modified SCSI driver.
- Net Dump
  - dump through the network by modified network card driver.
- LKCD (Linux* kernel crash dump)

- Kdump

  Easy to maintain.

  No low level hardware specific code.

  Dump method is more flexible.

  Require more system memory.

intel

Using kdump

Open Source Technology Center

- Build a kernel with CONFIG_KEXEC, CONFIG_CRASH_DUMP, CONFIG_PROC_VMCORE, CONFIG_DEBUG_INFO.

  We usually use the same kernel for dump capture kernel in IA-64.

- Boot this kernel with command line crashkernel="XXX@YYY", means reserve XXX amount of memory for dump capture kernel. On IA-64, the YYY usually should be 0, means kernel will choose the base address for dump capture kernel.

- Load crash dump kernel with "kexec –p <kernel-image> –initrd=<initrd> –append=<command line> maxcpus=1 irqpoll".

intel

Kdump

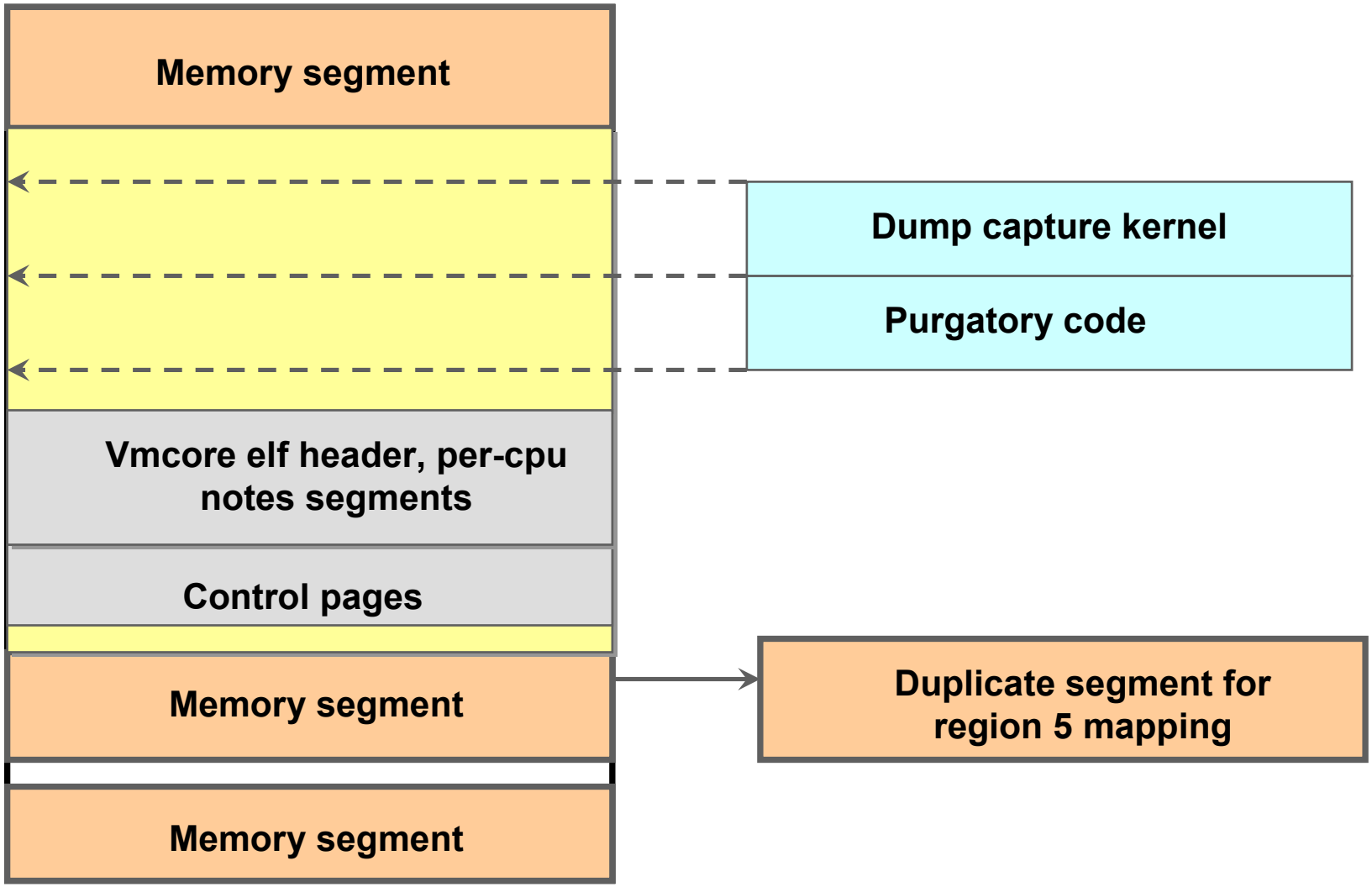
Open Source Technology Center

- Crash dump kernel region shows in /proc/iomem
- Trigger a crash, echo c > /proc/sysrq-trigger, or press INIT trigger on machine front panel.
- When the second kernel boots, copy /proc/vmcore to disk or do partial dump.
- Analyze crash dump with gdb <vmcore-file> <kernel-image>
- Or use crash utility to analyze dump result.

intel

Kdump

Open Source Technology Center



intel

Kdump

Open Source Technology Center

- Kdump hooks panic(). panic() will invoke crash_exec() if crashdump kernel is loaded.
- If kernel is not relocatable, kdump requires a build option to change physical address of kernel.
- Kdump needs to shutdown devices by himself, can't rely on driver->shutdown.
- Kdump needs to shoot down other CPUs himself.
- Dump capture kernel needs to know system memory layout of crashed kernel.

intel

Kdump on IA-64

Open Source Technology Center

- Reserve crash dump region according to size in command line
- IA-64 Linux* kernel is relocatable.
- Shutdown devices by send EOI then mask iosapic enties.
- Shoot down other CPUs by IPI+INIT message, CPU may loop or be send back to SAL rendez state.
- Collect system memory layout information when dump capture kernel is load into elf headers segment header.
- Save per-cpu registers in per-cpu notes segment , add an switch stack on top of current stack for crash utility to unwind.
- Limit dump capture kernel region by “max_addr= and min_addr=” in command line

intel

Kdump on IA-64

Open Source Technology Center

- INIT and unrecoverable MCA may be hooked by kdump.
- Purgatory code may split efi memmap entries and change the type of the entries if necessary.
- /proc/vmcore will be organized according to pre-stored elf header and segment header information.
- Partial dump may extract information from vmcore on system with huge amount of memory.
- Gdb or crash utility analyze crash according to dump file.

intel

Issues remain

Open Source Technology Center

- Some devices may not be shutdown properly which may cause problem on second kernel start.
- Need arch=dig command line option on platforms with IOMMU.
- For a platform without IOMMU, if there is no enough contiguous memory under 4G, the second kernel will not be able to use some devices.

intel

TODO

Open Source Technology Center

- More test for Kdump on different platforms
- Feedback issues, improvement suggestions.
- Some other possible usage model based on Kexec, e.g. a bootloader.

intel